# Hungarian Dependency Treebank

**Veronika Vincze[1], Dóra Szauter[1], Attila Almási[1], György Móra[1], Zoltán Alexin[2], János Csirik[3]**

[1]University of Szeged, Department of Informatics
Árpád tér 2., 6720 Szeged, Hungary
[2]University of Szeged, Department of Software Development
Árpád tér 2., 6720 Szeged, Hungary
[3]MTA-SZTE Research Group on Artificial Intelligence
Tisza Lajos krt. 103., III. lph., 6720 Szeged, Hungary

E-mail: {vinczev, szauter, gymora, alexin, csirik}@inf.u-szeged.hu, vizipal@gmail.com

**Abstract**

Herein, we present the process of developing the first Hungarian Dependency TreeBank. First, short references are made to dependency grammars we considered important in the development of our Treebank. Second, mention is made of existing dependency corpora for other languages. Third, we present the steps of converting the Szeged Treebank into dependency-tree format: from the originally phrase-structured treebank, we produced dependency trees by automatic conversion, checked and corrected them thereby creating the first manually annotated dependency corpus for Hungarian. We also go into detail about the two major sets of problems, i.e. coordination and predicative nouns and adjectives. Fourth, we give statistics on the treebank: by now, we have completed the annotation of business news, newspaper articles, legal texts and texts in informatics, at the same time, we are planning to convert the entire corpus into dependency tree format. Finally, we give some hints on the applicability of the system: the present database may be utilized – among others – in information extraction and machine translation as well.

## 1. Introduction

By converting the Szeged Treebank into syntactically annotated dependency trees, we aimed at creating the first manually annotated dependency corpus for Hungarian. The database may be utilized in various ways since besides its applicability in machine translation, it may function as a learning database in a number of information extraction systems. In this paper, we outline the corpus building process, present the problems and solutions, moreover, provide data on corpus statistics and finally give some hints on the applicability of the corpus and show how the database fits into international context.

## 2. Dependency grammars

Originally, in the Szeged Treebank, syntactic relations between sentence constituents are encoded in phrase-structured format. In the phrase-structured corpus, sentences are represented in a hierarchical structure made up of clauses: sentence constituents are organized into constituent trees. Clauses can be broken down into verbs, verbs can be broken down into arguments (nominal phrases) and other constituents, which, however, do not form a hierarchy below the NP level. The words of the sentence are located on the leaves of the constituent tree, the other nodes represent abstract units of organization (labeled with phrase-structure tags).

The dependency tree format differs from the constituent tree format inasmuch as every node in the tree corresponds to a word in the sentence. On the top of the sentence tree a virtual root node can be found to which words in the sentence are subordinated, that is, no abstract nodes can be found apart from the root node. Every word in the sentence is strictly subordinated to another one: a word can only have one superordinate, however, there can be several words below a node, e.g. all the arguments of a verb fall under the verb node. Nodes in the dependency tree can have diverse relations, usually tagged to denote the nature of the particular relation.

Tesnière's book (Tesnière, 1959) is considered to be the first dependency grammar, which lays the foundations of the theory. According to his famous metaphor, the verb is the central element of the sentence, which "*expresses a whole little drama*": the arguments of the verb are the actors, which Tesnière calls actants. Consequently, in a sentence subordinated and superordinated elements are integrated into a unit.

Mel'čuk's (1988; 2003) dependency grammar emerged within the Meaning-Text Theory. In this framework, dependency appears as a linear relation between words. On the deep syntactic level, he assumes twelve relation types, out of which six exist between the verb and its various arguments (actants) and the other relations designate coordination and diverse modifying roles. The heart of Mel'čuk's dependency grammar is that it interprets coordination as a kind of subordination: the conjunction is connected to the first member of coordination and the other member(s) of the coordination are connected to the latter with a special (COORD) relation. Another peculiarity of this approach is that in certain cases this grammar permits the insertion of nodes denoting abstract, that is, phonetically non-overt linguistic elements into the dependency tree: such is the case with the copula in Russian (and in Hungarian as well) in third person singular, present tense, which does not become overt in the sentence phonetically still it is there on an abstract level since it becomes manifest in past and

future tenses.
Koutny and Wacha (1991) and Prószéky et al. (1989) give a summary of a dependency grammar for Hungarian and the authors briefly outline their morpheme-based dependency grammar. In their model, morphemes are the basic constituents of dependency trees since in agglutinative languages not (only) words but morphemes too are capable of expressing different grammatical relations. This solution facilitates mapping between dependency trees of different types of languages because the node of e.g. the auxiliary *may* in English corresponds to the node of the morpheme *–hAt* in the Hungarian tree. This procedure may greatly enhance the efficiency of dependency grammar-based translation systems.

## 3. Dependency corpora for other languages

Dependency corpora have been developed for a number of languages. Among them, one of the most famous is the Prague Dependency Treebank developed for Czech (Hajič et al., 2000), which includes morphological, syntactical and tectogrammatical annotation. The same group has developed a dependency annotated parallel corpus for English and Czech (Čmerjek et al., 2004a, 2004b) and a dependency corpus for Arabic (Hajič et al., 2004). In addition to the above, dependency treebanks have already been developed for numerous European (e.g. Swedish (Nivre, 2003), Greek (Prokopidis et al., 2005), Russian (Boguslavsky et al., 2000), and Slovenian (Džeroski et al., 2006)) and non-European (Japanese (Lepage et al., 1998), Chinese (Liu, 2007)) languages and even for dead languages: a corpus for Latin has already been built and its authors are currently working on an Ancient Greek corpus (Bamman & Crane, 2006). By developing the first dependency corpus for Hungarian we wish to join this trend.

## 4. The corpus building process

In order to be able to convert the originally phrase-structured Treebank into a dependency corpus, first of all, a conversion step is needed during which constituent trees are converted into dependency relations. As automatic machine conversion is not expected to produce perfect, flawless results, this phase was followed by manual control, when linguists checked the files and modified them if necessary.
Although we can find a brief outline of the dependency grammar applied for Hungarian in the earlier literature (Koutny & Wacha, 1991; Prószéky et al., 1989), we did not follow this model completely when converting the Szeged Treebank into dependency tree format because this model is a morpheme-based one, that is, it is morphemes that are represented in the nodes of the dependency trees and not word forms. However, in order to be able to build syntactic trees from morphemes, we need a well-functioning morphologic parser capable of breaking down word forms in the Szeged Treebank into morphemes. Since there are no MSD-codes in the Szeged Treebank for derivation, the system treats causative suffixes and suffixes expressing possibility and permission (the suffix *–hat*/*het*) as part of the stem, thus, it would not be able to assign a separate morpheme, that is, a separate node to the suffix. Conversion into morpheme-based dependency trees would entail further, labor-intensive tasks (e.g. the transformation of the MSD-code system in such a way that derivation can be represented, the recoding of word forms within the corpus, the development of a well-functioning morphologic parser for the corpus etc.). For this reason, we undertook to mark dependency relations between word forms only.
The shared task announced by the organizing committee of CoNLL 2007 is considered as the first step of the conversion of the Szeged Treebank 2.0 into dependency tree format (Nivre et al., 2007). The conversion of the subcorpora containing articles from HVG and Népszabadság had been completed (Alexin, 2007), which process was later extended to the entire corpus.
In Szeged Treebank 2.0 linguistic relations between the verb and its arguments were marked. These relations had to be converted into dependency relations. For instance, instead of encoding all the twenty grammatical cases used for nominal complements in the constituency trees, only the cases nominative (SUBJ), accusative (OBJ), dative (DAT) were preserved and all the other cases were replaced by the tag OBL (obliquus). This unification of tags can be supported by the fact that since it is the former three grammatical cases that determine the basic arguments of a verb (i.e. subject, direct object, indirect object), and from an applicational viewpoint (for instance, in information extraction) it is usually sufficient to make a distinction between these and other arguments or adjuncts of the verb.
Retagging of the relations was done automatically on the basis of rules previously determined by linguists. Possible dependency relations are the following:

APPEND – non-integral parts of sentences
ATT – relation between noun and adjective, postposition and noun, noun/nominal modifier and noun
AUX – relation between verb and auxiliary
AUXS – node representing the whole sentence
CONJ – conjunction
COORD – coordination
DAT – dative (suffix *-nAk*)
DET – relation between noun and determiner
FROM – adverb or postpositional phrase answering for the question „from where?”
INF – infinitive
LOCY – adverb or postpositional phrase answering for the question „where?”
MODE – other adverbs or postpositional phrases
NEG – negative
OBJ – relation between verb and object
OBL – relation between verb and its other nominal argument
PRED – relation between verb and nominal predicate
PREVERB – relation between verb and preverb
PUNCT – punctuation mark
QUE – question word

ROOT – main element of the sentence
SUBJ – relation between verb and subject
TFROM – adverb or postpositional phrase answering for the question „from when?"
TLOCY – adverb or postpositional phrase answering for the question „when?"
TO – adverb or postpositional phrase answering for the question „where to?"
TTO – adverb or postpositional phrase answering for the question „till when or by when?"

To check the quality of the automatic conversion, we later compared the automatically converted and the manually annotated versions of two subcorpora (namely, newspaper texts from Népszava and Magyar Hírlap). The agreement rates were 63.246% and 62%, respectively, which underlines the necessity of manual annotation and correction.

### 4.1 Typical errors

Data yielded by the automatic conversion have been manually checked and corrected (if necessary) by four linguists. Errors fell into two typical categories: (1) the node was put at the wrong place in the tree; (2) the relation type between the node and its superordinate was not appropriate.
The majority of errors were due to the fact that not all linguistic relations were marked in the phrase-structured corpus, e.g. articles, numerals and attributes were included within the nominal phrase and their relation to the noun was not indicated. During automatic conversion, all these elements were linked to the noun with ATT relation and the other elements in the sentence to the verb with MODE relation. These, if it was necessary, had to be replaced with the right type dependency relation and to be removed to the appropriate superordinate (mother)-node.
The most frequent cases of retagging are the following:

- noun with a suffix within an attributive phrase

The converter – due to the above-mentioned reason – tagged every noun with ATT which was member in an AP (adjectival phrase), e.g. in *a ténylegesnél 1,9_milliárd dollárral magasabb árbevételt* 'the return 1.9 billion dollar more than the actual', *ténylegesnél* actual-ADE 'than the actual' and *1,9_milliárd dollárral* 1.9 billion dollar-INS '1.9 billion dollar' were tagged with ATT instead of the right OBL, so it had to be corrected.

- NEs

Named Entities were, in most cases, tagged with ATT, which had to be corrected according to the context.

- the tag of the main element in subordinate clauses

In the Treebank, demonstrative pronouns referring to the subordinate clause[1] were tagged in accordance with their roles within the main clause (and the subordinate clause was linked to the demonstrative pronoun if there was one present in the sentence). In the dependency corpus, however, we only indicated that it was a case of subordination, that is, we tagged the main element in the sentence with ATT.

- the second, third, … element in coordinations

In the Treebank, coordinations were labeled with an extra NP tag, in accordance with the usual solution in phrase-structure grammars, whose type agreed with the tag of the elements of coordination: thus, two coordinated noun phrases (NP) also had an external NP tag, which included both of them. As there are no artificial nodes in dependency grammars, this procedure could not be used, so we had to follow Mel'čuk's solution (1988; 2003) for the analysis of coordinations, see below.

- ez/az (*this/that*) determiners

Determiners were tagged with ATT if they occurred in a determiner + article + noun construction (*ez a ház* this the house 'this house'). When they occurred in nominative, they were tagged with DET, a tag for determiners and if they had a case ending (e.g. *ebben a házban* this-INE the house-INE 'in this house'), the tag had to be replaced with the right one for the particular case (for OBL in the present case).

Removal of nodes in the tree was most necessary in the cases presented below:

- subordinate clauses

Conjunctions did not form an integral part of subordinate clauses in the phrase-structured Szeged Treebank. As a result of this, after conversion both the conjunction and the main element in the subordinate clause were (severally) linked to the central element (root) in the main clause. During manual control, linguists linked the main element of the subordinate clause to the conjunction in this way establishing contact between the two components.

- possessive constructions

Often, the two parts of the possessive constructions, the possessor and the possessed, were not linked in the corpus. This especially applied to the possessor with the suffix –nAk[2], chiefly if it was not adjacent to the possessed in the sentence. In the dependency corpus, we always linked the possessor to the possessed, even if it produced "cross-dependencies", that is, if two edges in the tree intersected one another. (This is strictly forbidden in phrase-structure

[1] In Hungarian, a demonstrative pronoun referring to the subordinate clause can stand in the main clause as in *Azt mondta, hogy eljön* that-ACC say-PAST-3SG-DEF that come-PRESENT-3SG-INDEF 'He said that he would come'.

[2] In Hungarian, the possessor in possessive constructions can manifest in two forms: without any suffix and with the suffix *–nAk*.

grammars where movement is permissible, however, in dependency grammars intersection is accepted.)

- coordination

As has been mentioned in relation to the retagging cases, in coordination not only the tags of nodes but their position also had to be modified. During automatic analysis, the conjunction generally functioned as the main element of the construction and the members of coordination were in dependency relation with it. However, in accordance with Mel'čuk's (1988; 2003) solution, the first member of the construction functions as the main element and the conjunction (if there is any) has to be linked to it with CONJ relation, then follow the other members of coordination linked to the preceding element with COORD relation.

- infinitives and preverbs

If there was an (auxiliary) verb in the sentence that had an infinitival argument (*szeret* 'likes', *kíván* 'wishes', *fog* 'will', *kell* 'must'…), then automatic analysis linked the incidental preverb of the infinitive to the main verb. This type of error has also been corrected manually.

### 4.2 Coordination

Coordination poses problems for most theories of syntax: proponents of certain theories think it proper that the conjunction is the main element of the coordination, others argue that the head of the construction is a member of the coordination. Let us now examine these theories one after the other.

Let us postulate that the conjunction is the main element of the construction. The question arises what can be done in the cases of direct coordination when there is no conjunction between the elements. If there is no conjunction, we must postulate a virtual node capable of functioning as the main element. This theory, however, has another disadvantage: if there are more than two coordinated elements, then it is not possible to distinguish type "A and B and C" from type "A, B and C". This problem can be evaded in such a way that we assume an abstract "and" above "A" and "B", but then "B" would be linked to two nodes simultaneously (to a virtual AND and a real *and*) and this is strictly forbidden. A further disadvantage of this theory is that if e.g. the subject of a sentence is a coordinated phrase, then the verb and the conjunction would be linked with SUBJ relation and this is quite unusual.

According to another theory, coordinated elements and the conjunction are represented on the same level but they are not connected, e.g. in the construction *Jancsi és Juliska mézeskalácsháza* 'the candy-house of Hansel and Gretel', the following relations can be found: *mézeskalácsháza – Jancsi* 'candy-house' – 'Hansel' *mézeskalácsháza – és*, 'candy-house' – 'and' and *mézeskalácsháza – Juliska* 'candy-house' – 'Gretel'. In this case the problem is that though the relatedness of *Jancsi* and *Juliska* could somehow be indicated (with the ATT relation), however, it is problematic what relation to suppose between *mézeskalácsháza* and *és*, not to mention that it is quite unusual to leave the two members of the coordination unconnected.

Neither of the above approaches solve the problem satisfactorily, so we decided to follow Mel'čuk's theory of coordination (1988; 2003) where coordination is treated as kind of "subordination". The main element is always the first member in the coordination because it is capable of functioning as an entire phrase. Let us now examine the following example:

*Elmentem a boltba Józsival és Kival*.
'I went shopping with Joe and Katie.'
*Elmentem a boltba Józsival.*
'I went shopping with Joe.'
**Elmentem a boltba Józsival és.*
'*I went shopping with Joe and.'
**Elmentem a boltba és Kival*.
'*I went shopping and Katie.'

The difference between the second, the third and fourth sentences show that coordination cannot be split into two equal parts since if the elements *Józsival* 'with Joe' and *és Kival* 'and Katie' were equivalent, then the last sentence should be acceptable. *Józsival* 'with Joe' is not closely connected to *és* 'and' either since in that case the third sentence would also be grammatical. The solution is that we postulate three parts in the coordination: the first is the main element, the conjunction is linked to it with CONJ relation and the conjunction is followed by the second coordinated member with COORD relation as in Figure 1. From a representational perspective this is in fact subordination and so there will be no difference of structure between coordination and subordination: only the relations (COORD and ATT, respectively) indicate which is which.

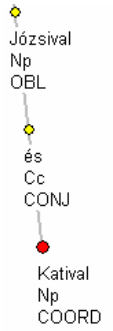


Figure 1: Coordination

### 4.3 Predicative nouns and adjectives

Owing to the peculiarities of Hungarian language, in sentences containing predicative nouns or adjectives, the declarative, third person singular, present tense form of the copula does not become overt as opposed to forms in different mood, tense, number and person:

*András katona (*van).*
'Andrew is a soldier.'
*András legyen katona.*
'Let Andrew be a soldier.'
*András katona lesz.*
'Andrew is going to be a soldier.'

Similarly to coordination, there are two ways to solve this particular problem.
First, the main element of the sentence is the predicative noun (or adjective); the subject is linked to it and no virtual node is assumed. The disadvantage of this solution is that a completely different structure is ascribed to the same sentence in the present third person singular (see Figure 2) and all the other tenses / persons (see Figure 3), which is questionable because in one case there is direct, while in the other case there is indirect relation between the predicative element and the subject.

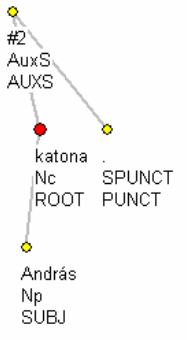


Figure 2: A sentence with a non-overt copula

Compare Figures 2 and 3:

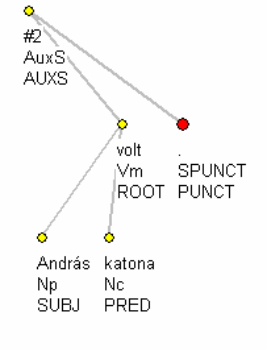


Figure 3: A sentence with an overt copula

Second, the same structure is retained for any occurrence of the sentence, it is true, however, that the price of it is that a virtual node has to be postulated for the declarative, third person singular, present tense form of the copula (VAN). In this way, dependency trees are structured as in Figure 4.
A further argument for the use of a virtual node is that VAN is by all means present on the syntactic level since it is overt in all the other forms, tenses and moods of the verb. It is only a secondary (morphological) question why its third person singular, present tense form is a zero morpheme (cf. Mel'čuk, 2003). The use of virtual nodes may have advantages with regard to the international applicability of the corpus since e.g. a translator program based on dependency trees is a lot more effective if it is to map a tree with similar structure to another language as opposed to that if even extra transformational steps have to be inserted into the translation process.

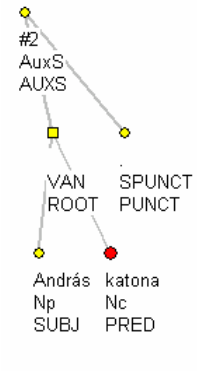


Figure 4: A sentence with a virtual copula

## 5. Statistics

The file of the Szeged Treebank 2.0 contains 82.000 sentences, 1.2 million words and 250.000 punctuation marks. Texts were selected from six different domains, ~200.000 words in size from each. The domains are the following:

- fiction
- compositions of pupils between 14-16 years of age
- newspaper articles (from the newspapers Népszabadság, Népszava, Magyar Hírlap, HVG)
- texts in informatics
- legal texts
- business and financial news

The format of the database follows the CoNLL 2007 Shared Task norms (Nivre et al., 2007). It is freely available for research and educational purposes at http://www.inf.u-szeged.hu/rgai.
Statistical data on the so far completed corpus are represented in the charts below.

| | newsml | newspapers | law | informatics | total |
|---|---|---|---|---|---|
| sentences | 9574 | 10210 | 9278 | 9759 | 38821 |
| words | 186030 | 182172 | 220069 | 175991 | 764262 |
| punctuation marks | 25712 | 32880 | 33515 | 31577 | 125622 |

Table 1: Number of sentences, words and punctuation marks in the corpus

The most frequent dependency relations occurring in the

corpus can be seen in Table 2. The most frequent one, ATT is a general 'modifier' relation encoding attributive and subordination relations, that is, it can relate words and clauses as well. Maybe, its high frequency in the corpus can be attributed to the above-mentioned fact. Since nouns usually occur together with some kind of determiner, the relation DET can be expected to occur quite often as well. The third most frequent relation, OBL is a superordinate relation of several cases in Hungarian declination that is why its frequency is based on the sum of the frequency of such cases.

| | Relation type | Percentage |
|---|---|---|
| 1. | ATT | 32.3% |
| 2. | DET | 14.8% |
| 3. | OBL | 11% |
| 4. | SUBJ | 7% |
| 5. | CONJ | 6.3% |
| 6. | COORD | 5.4% |
| 7. | MODE | 5.3% |
| 8. | OBJ | 4.7% |
| 9. | ROOT | 4.4% |
| 10. | TLOCY | 1.8% |

Table 2: The most frequent relation types

Corpus texts are annotated by four linguists. They have regular meetings where recurring annotation problems are discussed and solved. At the beginning of the annotation project, 700 sentences were annotated by all the four linguists on the basis of which agreement rates were calculated. The agreement rates between the annotators are listed in Table 3:

| | A#1 | A#2 | A#3 | A#4 |
|---|---|---|---|---|
| A#1 | - | 92.94% | 92.73% | 94.97% |
| A#2 | 92.94% | - | 91.81% | 94.06% |
| A#3 | 92.73% | 91.81% | - | 93.64% |
| A#4 | 94.97% | 94.06% | 93.64% | - |

Table 3: Agreement rates between annotators

The overall agreement rate is 93.36%.

## 6. The applicability of the corpus

Applying dependency trees has advantages in several fields of computational linguistics: corpora in dependency-tree format may be used successfully in both machine translation and information extraction.

### 6.1 Machine translation

Machine translation processes based on syntactic transformation rely on two methods: they either map the constituent trees of the source language to the constituent trees of the target language or work with dependency trees. One of the advantages of the method using constituent trees is that it may very well be used for machine translation of cognate languages since the syntax of these languages is usually similar, moreover it sufficiently solves the problem arising from differences in word order. Its disadvantage is that complicated and costly transformation rules have to be introduced to the system, furthermore, if the sentence has a completely different syntactic structure in the source and the target language, automatic translation becomes totally unacceptable.

Another common error in the translation systems using constituent trees is that the parser often ascribes incorrect structure to the tree, inserts redundant, unnecessary tags or matches nodes wrongly. Dependency tree-based translation systems successfully eliminate the errors arising from virtual nodes as there are no abstract (virtual) nodes in dependency trees. Each node in the tree corresponds to a natural language element, the tree does not contain syntactic nodes so the syntactic differences disappear. In the machine translation process every node gets translated and if necessary, nodes reorganize along previously given probabilities. The machine translation process using dependency trees is especially rewarding in the case of non-cognate languages or language pairs with different syntax.

### 6.2 Information extraction

Dependency trees can be used in another field of computational linguistics, i.e. in information extraction. Syntactically annotated corpora play an important role in automatic information extraction for it is not enough to know what words and expressions are included in the given text, their relation is of great significance as well. For instance, in business texts, it must be included in the information on different transactions that if company A and B took part in a business transaction, which company bought up the other (that is, which company is the subject and which is the object of the verb *buy up*). However, in order to be able to make the right decision, the information extraction system has to be capable of parsing syntactic relations as well. Syntactically annotated corpora have a great part in training syntactic parsers to analyze relations.

In the case of languages with fixed word-order a syntactically annotated corpus using constituent trees is a good alternative for in these corpora a given syntactic structure is associated with a given syntactic relation. Dependency grammar-based corpora, however, are of great help in information extraction in the case of free word-order languages since in their case word-order is of no use with respect to syntactic relations: the gist of dependency grammars is that they are capable of identifying the syntactic structure of the sentence irrespective of word-order.

Basic relations between the verb and its argument are indicated in the present corpus, that is the subject, object and dative arguments can be identified easily (tagged with SUBJ, OBJ and DAT, respectively) and the other arguments have OBL tags. In this way, the information extraction program can successfully identify the syntactic relations in the following example:

*Az E.ON_Hungária_Energetikai_Rt. 87,713 százalékra növelte részesedését a*

*Tiszántúli_Áramszolgáltató_Rt-ben.*

'E.ON Hungária Energetikai Rt. increased its share in Tiszántúli Áramszolgáltató Rt. to 87.713 percent.'

The relevant syntactic relations to be extracted are the following:

*növelte - Az E.ON_Hungária_Energetikai_Rt.*
'increased' – 'E.ON Hungária Energetikai Rt.' (subject)
*növelte – részesedését*
'increased' – 'its share' (object)
*növelte – a Tiszántúli_Áramszolgáltató_Rt-ben*
'increased' – 'in Tiszántúli Áramszolgáltató Rt.' (argument)

From the syntactic relations it becomes clear even for the computer what relation the two Named Entities in the sentence have, that is, it is E.ON that has a share in Tiszántúli Áramszolgáltató Rt. and not vice versa. In this way the precision of information extraction using syntactic relations improves greatly as compared with models not using them.

### 6.3 Multilinguality

The development of the Hungarian dependency corpus opens the door to multilingual applications. *1984* and the *Windows2000* text files in the subcorpora of the Szeged Treebank may be the link to multilingual parallel corpora since these texts surely have a parallel in a foreign language. If the foreign language versions contain syntactic annotation based on dependency relations, a parallel dependency corpus for Hungarian and the given foreign language can easily be produced. This would greatly help with the development of – on the one hand – systems supporting multilingual information extraction and – on the other hand – translation programs using syntactic methods. Therefore, building such a corpus can be considered a significant and hopeful effort from both theoretical and practical perspectives.

## 7. Summary

In this paper, we have presented the process of the conversion of the Szeged Treebank into dependency tree format: delineated the work process, presented the problems and their solutions, demonstrated its applicability in machine translation and information extraction, moreover, showed its advantages for researchers in contrastive linguistics and dependency syntax. Further on, we would also like to implement a dependency parser for Hungarian (possibly by adapting an already available one (e.g. the MaltParser (Nivre et al., 2007)) or by developing our own), for which this corpus can be used as a learning database.

## 8. Acknowledgements

The research was – in part – supported by NKTH within the framework of TUDORKA and MASZEKER projects (Ányos Jedlik programs).

## 9. References

Alexin, Z. (2007). A frázisstrukturált Szeged Treebank átalakítása függőségi fa formátumra [Converting the phrase-structured Szeged Treebank into dependency format]. In Tanács, A., Csendes, D. (Eds.), *V. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2007)*. Szeged: University of Szeged, pp. 263--266.

Bamman, D., Crane, G. (2006). The Design and Use of a Latin Dependency Treebank. In *Proceedings of the Fifth International Workshop on Treebanks and Linguistic Theories (TLT 2006) (Prague)*, pp. 67--78.

Boguslavsky, I., Grigorieva, S., Grigoriev, N., Kreidlin, L., Frid, N. (2000). Dependency Treebank for Russian: Concept, Tools, Types of Information. In *Proceedings of the 18th Conference on Computational linguistics*. Saarbrücken, Germany, pp. 987--991.

Čmejrek, M., Cuřín, J., Havelka, J. (2004). Prague Czech-English Dependecy Treebank: Any Hopes for a Common Annotation Scheme? In *HLT/NAACL 2004 Workshop: Frontiers in Corpus Annotation*. Boston, Massachusetts, pp. 47--54.

Čmejrek, M., Cuřín, J., Havelka, J., Hajič, J., Kuboň, V. (2004). Prague Czech-English Dependecy Treebank: Syntactically Annotated Resources for Machine Translation. In *4th International Conference on Language Resources and Evaluation*. Lisbon, Portugal.

Csendes D., Csirik J., Gyimóthy T., Kocsor A. (2005). The Szeged Treebank. In *Proceedings of the Eighth International Conference on Text, Speech and Dialogue (TSD 2005)*. Karlovy Vary, Czech Republic and LNAI series Vol. 3658, pp. 123--131.

Džeroski, S., Erjavec, T., Ledinek, N., Pajas, P., Žabokrtský, Z., Žele, A. (*2006). Towards a Slovene Dependency Treebank.* In *Proceedings of Fifth International Conference on Language Resources and Evaluation, LREC'06*. Genoa, Italy.

Hajič, J., Böhmová, A., Hajičová, E., Vidová Hladká, B. (2000). The Prague Dependency Treebank: A Three-Level Annotation Scenario. In A. Abeillé (Ed.), *Treebanks: Building and Using Parsed Corpora*. Amsterdam: Kluwer, pp. 103--127.

Hajič, J., Smrž, O., Zemánek, P:, Šnaidauf, J., Beška, E. (2004). Prague Arabic Dependency Treebank: Development in Data and Tools. In *Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools*. Cairo, Egypt, pp. 110--117.

Koutny, I., Wacha, B. (1991). Magyar nyelvtan függőségi alapon [A dependency-based grammar of Hungarian]. *Magyar Nyelv*, 87(4), pp. 393--404.

Lepage, Y., Shin-Ichi, A., Susumu, A., Hitoshi, I. (1998). An annotated corpus in Japanese using Tesnière's structural syntax. In *Proceedings of COLING-ACL'98 Workshop on the Processing of Dependency-based Grammars*. Montreal.

Liu, H. (2007). Building and Using a Chinese Dependency Treebank. *GrKG/Humankybernetik* 48(1), pp. 3--14.

Mel'čuk, I. A. (1988). *Dependency Syntax: theory and practice*. Albany, NY: State University of New York Press.

Mel'čuk, I. A. (2003). Levels of Dependency in Linguistic Description: Concepts and Problems. In Agel, V., Eichinnger, L., Eroms, H.-W., Hellwig, P.,

Herringer, H. J., Lobin, H. (Eds.), *Dependency and Valency. An International Handbook of Contemporary Research*. Vol. 1. Berlin-New York: W. de Gruyter, pp. 188--229.

Nivre, J. (2003). Theory-Supporting Treebanks. In Nivre, J., Hinrichs, E. (Eds.), *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*. Växjö University Press, pp. 117--128.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., Yuret, D. (2007). The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*. Prague, pp. 915--932.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., Marsi, E. (2007). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2), pp. 95--135.

Prokopidis, P., Desipri, E., Koutsombogera, M., Papageorgiou, H., Piperidis, S. (2005). Theoretical and practical issues in the Construction of a Greek Dependency Corpus. In *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT-2005)*. Barcelona.

Prószéky, G., Koutny, I., Wacha, B. (1989). Dependency Syntax of Hungarian. In Maxwell, D., Schubert, K. (Eds.), *Metataxis in Practice (Dependency Syntax for Multilingual Machine Translation)*. Dordrecht, The Netherlands: Foris, pp. 151--181.